AUTOMATED PROTEOME-WIDE DETERMINATION OF SUBCELLULAR LOCATION USING HIGH THROUGHPUT MICROSCOPY

Robert F. Murphy

Ray and Stephanie Lane Center for Computational Biology, Center for Bioimage Informatics, and Departments of Biological Sciences, Biomedical Engineering, and Machine Learning, Carnegie Mellon University, Pittsburgh PA

ABSTRACT

A major source of information for identifying subcellular location on a proteome-wide basis will be imaging of tagged proteins in living cells using fluorescence microscopy. We have previously developed automated systems to interpret images from such experiments and demonstrated that they can perform as well or better than visual inspection. Recent work demonstrates that these methods can be applied to large collections of images from sources as diverse as yeast expressing GFP-tagged proteins and human tissues imaged by immunocytochemistry. A distinct but related task is learning what location patterns exist. We have demonstrated clustering of mouse proteins into subcellular location families that share a statistically indistinguishable pattern. To communicate each pattern, we have developed approaches to learning generative models of subcellular patterns. Integration of high-throughput microscopy and automated model building with cell modeling systems will permit accurate, well-structured information on subcellular location to be incorporated into systems biology efforts.

Index Terms— Location proteomics, tissue micro-array, pattern recognition, generative models, high throughput microscopy

1. INTRODUCTION

An important challenge in the post-genomic era is to identify subcellular location on a proteome-wide basis. High-throughput microscopy systems provide an important capability to enable this task, especially when combined with tagging of proteins in living cells using fluorescence protein fusions. The large volume of images generated by high throughput systems requires automated systems for interpretation. Automated systems not only can recognize all major subcellular patterns [1-3], but they can perform as well or better than visual inspection [4-6]. Examples of major patterns used for development and testing of these systems are shown in Figure 1. Whether automated approaches can be applied to sets of proteins approaching the proteome size has not been clear. We discuss here approaches to comprehensively and systematically analyzing protein subcellular location and especially how the resulting knowledge can be integrated into predictive cell models.

2. PROTEOME-WIDE PATTERN CLASSIFICATION

Initial work on subcellular pattern analysis was focused on images of cultured cells for a small set of proteins known to localize to each of the major subcellular structures. An important question therefore was whether such methods could be extended to larger image collections and more difficult cellular contexts. The recent public availability of image collections for large numbers of proteins has made addressing that question feasible. An important example is the UCSF yeast GFP (green fluorescent protein) localization database, which contains images of GFP-fusions for most suspected protein-coding regions in *S. cerevisiae* [7]. Each

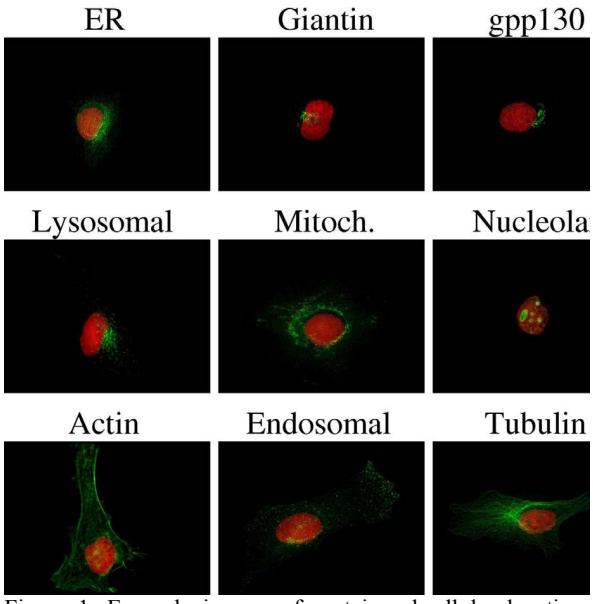

Figure 1. Example images of protein subcellular location patterns from the 2D HeLa collection [1] (available from http://murphylab.web.cmu.edu/data). DNA distributions are shown in red and protein distributions are shown in green.

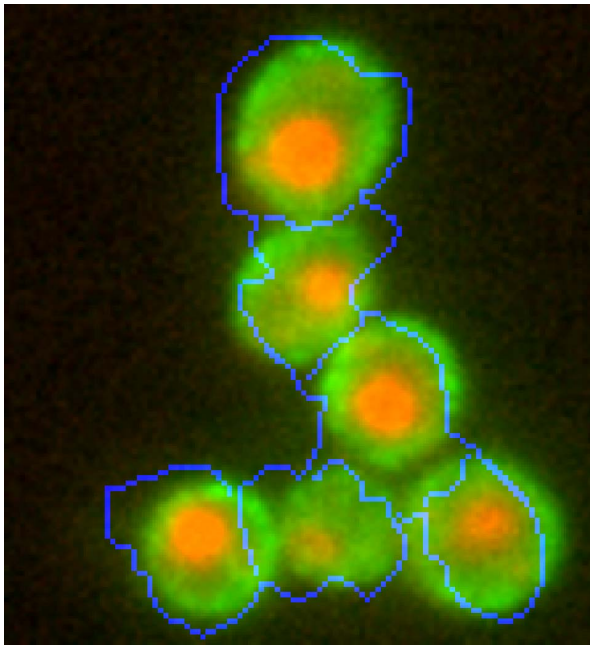

Figure 2. Portion of image of ORF YGR130C downloaded from the UCSF yeast GFP fusion localization database ((http://yeastgfp.ucsf.edu). The DNA distribution is shown in red, the estimated cell boundary found during cell segmentation is shown in blue, and the GFP-fusion protein distribution is shown in green. This protein was classified as a punctate_composite protein in the UCSF database and classified as a cell_periphery protein by automated localization with 60.7% confidence. The CYGD database annotates it as a mixture of cytoplasm and punctate_composite protein.

image in the collection was annotated by two human curators using one or more of 22 subcellular location terms. The difficulty of analyzing this collection stems from the small size of yeast cells (relative to mammalian cells for which all previous automated analysis has been done) and the presence of clumps of cells and out-of-focus cells in the images in the collection. Since common cell segmentation methods such as seeded watershed did not work well for this collection, we developed a graphical model-based method for segmenting the images and removing cells that did not show expected ellipsoidal geometry [8]. Using this method combined with the Subcellular Location Features we have described previously, we built classifiers for those images annotated as belonging to only one location class [9]. The accuracy of this classifier was over 80%, and that accuracy increased to nearly 95% when only proteins for which the classifier estimated a high confidence were considered. Interestingly, for the proteins for which the high-confidence assignments differ from the human annotations, re-examination of the images suggests that at least some of the automated assignments are more likely to be correct. An example image for a protein whose automated assignment appears to be more accurate than the human assignment is shown in Figure 2. Further work will be needed to resolve the differences between visual and automated assignments, but the approach described should be useful for automatically annotating subcellular location for new yeast species, for strains with different genotypes, or for a given strain under different conditions.

Another important publicly-available collection is the Human Protein Atlas, which contains images for thousands of proteins in all major human tissues [10]. These images were collected using immunocytochemistry with well-characterized mono-specific antibodies and an automated imaging platform, with an initial goal of documenting the level of expression of each protein in each tissue. While the images have lower resolution than those previously used for automated subcellular pattern analysis, we have recently obtained encouraging results demonstrating the feasibility of training a single classifier to recognize the major subcellular patterns across all tissue types [11]. These results set the stage for analyzing variation in subcellular pattern (if any) for each protein from tissue to tissue.

3. LEARNING SUBCELLULAR PATTERNS USING CLUSTER ANALYSIS: SUBCELLULAR LOCATION FAMILIES

The development of the systems mentioned above that are capable of assigning proteins to major subcellular location categories has been an important step in demonstrating the applicability of automated image analysis approaches to fluorescence microscope images. However, we have previously proposed that unsupervised methods are more appropriate to the analysis of protein subcellular location patterns [4]. We have used the retroviral CD-tagging technology developed by Jarvik, Berget and colleagues [12] to collect increasing numbers of images of mouse 3T3 cells expressing proteins randomly-tagged with GFP and then cluster them into Subcellular Location Trees [6, 13, 14]. As the number of tagged lines examined has increased, the number of statistically distinguishable clusters has also increased (Table 1). The number of clones examined is currently over 1,000 and growing (unpublished data).

Number of clones	Number of clusters found	Reference
46	12	[13]
87	17	[14]
126	35	[6]
174	41	[6]

Table 1. Estimating number of statistically distinguishable subcellular location patterns in 3T3 cells.

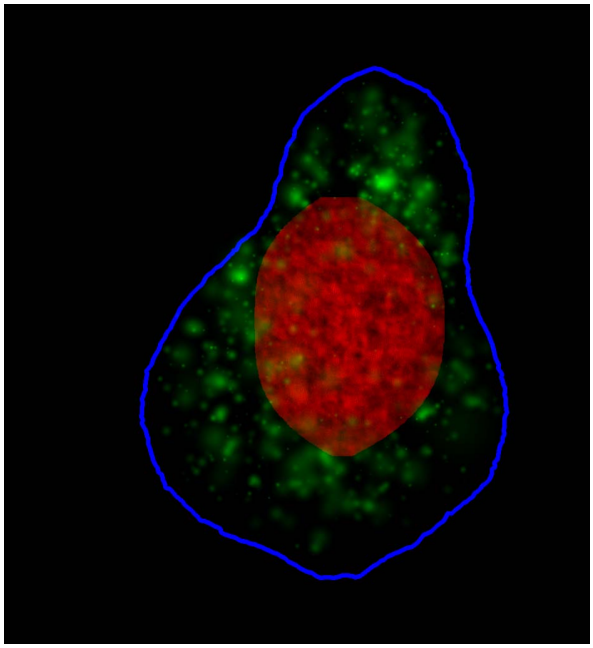

Figure 3. Example image synthesized from a generative model of an endosomal pattern. The model was trained on images of the distribution of transferrin receptor. The synthetic DNA distribution is shown in red, the plasma membrane boundary is shown in blue, and endosomes are shown in green. Synthetic images like this were recognized as endosomal with 91% accuracy by a machine classifier trained on real endosomal images [16].

This approach groups proteins that show patterns that are statistically indistinguishable (at least under the conditions used for imaging), and many of these proteins are likely to be part of stable complexes. A complementary approach is to search for unique combinations of proteins that are found within a single pixel or region using images obtained by repeated cycles of staining of fixed cells (an approach termed MELK) [15]. This identifies proteins which may interact but do not necessary remain together throughout the cell.

4. CAPTURING AND COMMUNICATING SUBCELLULAR LOCATION PATTERNS: GENERATIVE MODELS

The ability to group proteins into location families without human intervention has powerful implications for using high-throughput microscopy to characterize proteins on a proteome-wide basis. However, it begs the question of how to communicate what distinguishes each family in the absence of a priori category definitions. For this purpose, we have proposed that a generative model can be used to represent each family, much the same as generative Hidden Markov Models can be used to summarize sequence families. We have therefore developed approaches to directly learning generative models of subcellular patterns from images [16]. These can be used to synthesize images that in a statistical sense are drawn from the same underlying population as the images used for training. An example of a generated image for the endosomal (Transferrin Receptor) pattern is shown in Figure 3. The models can be communicated in compact XML files that are compatible with cell model descriptions captured in SBML. We anticipate combining these models to construct cell models containing all expressed proteins in their proper locations. We are currently working to integrate our tools with existing cell modelling systems, such as Virtual Cell [17] and MCell [18], to permit accurate, well-structured information on subcellular location to be incorporated into systems biology efforts.

5. ACKNOWLEDGMENTS

The work from my group summarized here was supported by in part by NSF ITR grant EF-0331657 and NIH grants GM068845 and GM75205. Facilities and infrastructure were supported by NIH grant U54 DA0215 (Dr. Brian Athey, PI) and by NIH grant U54 RR022241 (Dr. Alan Waggoner, PI).

6. REFERENCES

[1] M. V. Boland and R. F. Murphy, "A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells," *Bioinformatics*, vol. 17, pp. 1213-1223, 2001.

[2] C. Conrad, H. Erfle, P. Warnat, N. Daigle, T. Lorch, J. Ellenberg, R. Pepperkok, and R. Eils, "Automatic identification of subcellular phenotypes on human cell arrays," *Genome Research*, vol. 14, pp. 1130-1136, 2004.

[3] N. Hamilton, R. Pantelic, K. Hanson, and R. Teasdale, "Fast automated cell phenotype image classification," *BMC Bioinformatics*, vol. 8, pp. 110, 2007.

[4] R. F. Murphy, M. Velliste, and G. Porreca, "Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images," *J VLSI Sig Proc*, vol. 35, pp. 311-321, 2003.

[5] E. Glory and R. F. Murphy, "Automated Subcellular Location Determination and High Throughput Microscopy," *Developmental Cell* vol. 12, pp. 7-16, 2007.

[6] E. Garcia Osuna, J. Hua, N. Bateman, T. Zhao, P. Berget, and R. Murphy, "Large-Scale Automated Analysis of Location Patterns in Randomly Tagged 3T3 Cells," *Annals Biomed. Eng*, vol. 35, pp. 1081-1087, 2007.

[7] W.-K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea, "Global analysis of protein localization in budding yeast," *Nature*, vol. 425, pp. 686-691, 2003.

[8] S.-C. Chen, T. Zhao, G. J. Gordon, and R. F. Murphy, "A novel graphical model approach to segmenting cell images," *Proceedings of the IEEE Symposium on Computational*

Intelligence in Bioinformatics and Computational Biology, pp. 1-8, 2006.

[9] S.-C. Chen, T. Zhao, G. J. Gordon, and R. F. Murphy, "Automated Image Analysis of Protein Localization in Budding Yeast," *Bioinformatics* vol. 23, pp. i66-i71, 2007.

[10] M. Uhlen, E. Bjorling, C. Agaton, C. A.-K. Szigyarto, B. Amini, E. Andersen, A.-C. Andersson, P. Angelidou, A. Asplund, D. Cerjan, M. Ekstrom, A. Elobeid, and C. Eriksson, "A human protein atlas for normal and cancer tissues based on antibody proteomics," *Amer Soc Biochem Mol Biol*, vol. 4, pp. 1920-1932, 2005.

[11] J. Newberg and R. Murphy, "A Framework for the Automated Analysis of Subcellular Patterns in Human Protein Atlas Images," *J. Proteome Res.*, pp. in press, 2008.

[12] J. W. Jarvik, G. W. Fisher, C. Shi, L. Hennen, C. Hauser, S. Adler, and P. B. Berget, "In vivo functional proteomics: Mammalian genome annotation using CD-tagging," *BioTechniques*, vol. 33, pp. 852-867, 2002.

[13] X. Chen, M. Velliste, S. Weinstein, J. W. Jarvik, and R. F. Murphy, "Location proteomics - Building subcellular location trees from high resolution 3D fluorescence microscope images of randomly-tagged proteins," *Proceedings of SPIE*, vol. 4962, pp. 298-306, 2003.

[14] X. Chen and R. F. Murphy, "Objective clustering of proteins based on subcellular location patterns," *J Biomed Biotechnol*, vol. 2005, pp. 87-95, 2005.

[15] W. Schubert, B. Bonnekoh, A. J. Pmmer, L. Philipsen, R. Bockelmann, Y. Malykh, H. Gollnick, M. Friedenberger, M. Bode, and A. W. M. Dress, "Analyzing proteome topology and function by automated multi-dimensional fluorescence microscopy," *Nat Biotechnol*, vol. 24, pp. 1270-1278, 2006.

[16] T. Zhao and R. F. Murphy, "Automated learning of generative models for subcellular location: building blocks for systems biology," *Cytometry A*, vol. 71, pp. 978-90, 2007.

[17] Moraru, II, J. C. Schaff, B. M. Slepchenko, and L. M. Loew, "The virtual cell: an integrated modeling environment for experimental and computational cell biology," *Ann N Y Acad Sci*, vol. 971, pp. 595-6, 2002.

[18] J. S. Coggan, T. M. Bartol, E. Esquenazi, J. R. Stiles, S. Lamont, M. E. Martone, D. K. Berg, M. H. Ellisman, and T. J. Sejnowski, "Evidence for Ectopic Neurotransmission at a Neuronal synapse," *Science*, vol. 309, pp. 446-451, 2005.